Statistical Learning: Assignment 2

Sherry Usman

May 22, 2024

Question 1

The dataset for this experiment is derived from a large epidemiology study regarding the progression of depression and anxiety disorders among individuals living in the Netherlands.

In order to accurately assess the supervised learning methods needed for this dataset it is necessary to do an initial exploratory analysis.

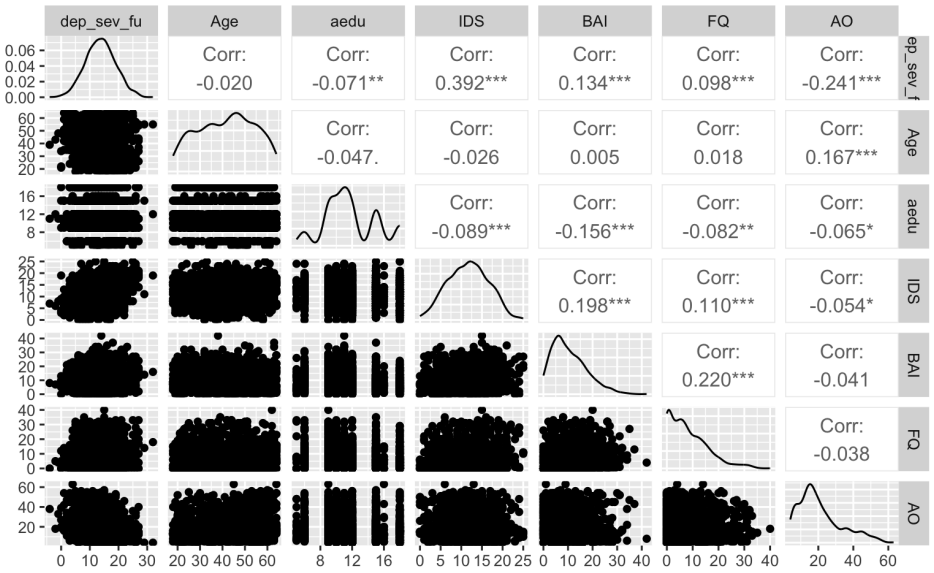

Figure 1: The figure shows an initial exploration of the dataset.

The figure above shows the distributions of variables in our dataset. A cursory glance shows that while variables such as *dep_sev_fu* and *IDS* are normally distributed, variables such as *BAI*, *FQ* and *AO* are more heavily positively skewed. A deeper correlation test using Pearson correlation (in Appendix 1) shows that the relationship between *dep_sev_fu* and other variables is mainly non-linear. Thus supervised learning methods such as linear regression would not work since they assume a linear relationship between the outcome and predictor variables. Thus, we explore other methods that can more accurately capture non-linear relationships between predictor and outcome variables. To accomplish this task we choose three supervised learning methods:

- Support Vector Regression (SVRs)
- Generalised Additive Models (GAMs)
- Gradient Boosting Machines (GBMs)

All three methods are known to capture non-linear relationships between predictor and outcome variables accurately.

In order to understand the choice of these particular supervised learning methods it is important to dive into their characteristics and why those characteristics fit our dataset well. We start with SVRs, the extension of the more commonly known Support Vector Classifiers (SVCs or SVMs). While SVMs attempt to classify different datapoints into classes by inserting a hyperplane that maximises the margin between different classes of datapoints, SVRs extend this to predict continuous outcomes in regression problems. These margins can be linear in the case of linearly separable data but can also be soft or non-linear in cases of data which is not clearly separable by a linear boundary. This flexibility to non-linear data makes SVR one of the contending methods for our dataset. Furthermore, SVRs also include a regularisation parameter C which prevents overfitting by controlling the bias-variance trade-off between obtaining the maximum margin and achieving minimum classification error. Lastly, SVRs provide a number of suitable parameters such as epsilon ϵ to vary their robustness to outliers in data. For all these reasons, SVRs seem to be a good choice for the *MH_dat* dataset.

Another supervised learning method that was looked into was GAMs. GAMs are a variation of generalised linear models which includes flexible functions for each predictor variable. These functions allow for flexible non-linearities in variables but still retains the additive structure of linear models. Thus, the relationships between predictor and outcome variables are not assumed to be linear, making them suitable for this case. This is shown in detail in the equation below.

$$y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + ... + f_p(x_{ip}) + \epsilon_i \tag{1}$$

where f_1, f_2 ... f_p are non-linear functions fitted to predictor variables x_{i1}, x_{i2} ... x_{ip}. GAMs while being simple and easy to use also allow us to play around with the smoothing functions and interact with the data on a closer and more transparent level. Another important advantage of using GAMs is that GAMs produce nice, easy-to-interpret plots of the contributions of different predictors, making it easy to see the positive or negative effects of each variable on the outcome.

The prospect of implementing a supervised learning method that avoided the strict smoothing assumptions of SVRs and GAMs was also exciting. Tree-based methods for instance are different from the methods listed above as they focus on stratifying or segmenting the predictor space into regions rather than fitting functions on existing data. While there are a myriad of different decision-tree methods available, the method that was chosen for this particular dataset was Gradient boosting.

Gradient boosting is a decision-tree ensemble method that combines the advantages of multiple models by sequentially fitting models. It does that by first fitting an initial regression tree to the data and then building subsequent trees that rectify the residual errors produced by the initial tree. This sequential approach helps to create a final model that has more refined predictions and a lower bias than its constituent models. Gradient boosting while computationally expensive and time-consuming can be much more powerful than other existing ensemble methods like bagging as its sequential approach reduces both bias and variance while other existing ensemble methods primarily reduce variance by averaging. That is why it is chosen for this dataset.

Question 2

To understand what parameters to choose it is first important to understand how each method works and what role each parameter plays in model performance.

SVM with Linear kernel: Since the dataset shows evidence of non-linearity it was decided implement an SVR with a linear kernel with softer margins and to implement SVRs with polynomial and radial-basis kernel.
The decision function for a linear kernel is defined as follows:

$$f(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \tag{2}$$

where x is the set of predictor variables, f(x) is the outcome variable, β_0 is the intercept term and β is the set of coefficients corresponding to the set of predictor variables. For an SVR with a linear

kernel with softer margins, the goal remains the same: to maximise the margin M. We can define this as follows:

$$maximise \text{ M subject to } \sum_{j=1}^{p} \beta_j^2 = 1, \tag{3}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ...\beta_p x_{ip} \geq M(1 - \epsilon_i), \tag{4}$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C \tag{5}$$

where ϵ is the margin of tolerance and C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the error. More precisely, a higher C indicates less slack (more penalty for errors) and a lower C indicates more slack (less penalty for errors).
Since we do not have an initial range of C and ϵ values that might be suitable, we cross-validate for parameters C and ϵ by exploring model performance over a broad range of values for C and ϵ and selecting the parameters that yield the lowest MSE or MAE.

SVM with Radial Basis kernel: For SVRs with a radial basis kernel, a radial basis function is used and is defined as follows:

$$K(x_i, x_i') = \exp(-\gamma \sum_{j=1}^{p} x_{ij} - x_{i'j})^2 \tag{6}$$

where x_i and x_i' are different input datapoints on the plane and γ is a hyperparameter that determines how far the influence of a single observation reaches. For example a lower gamma yields smoother decision boundary further reach and a higher gamma yields a more complex decision boundary with a smaller reach. Furthermore, there are additional parameters such as regularisation parameter C and *epsilon* incorporated in the objective function. The regularisation parameter C performs similar functions as the one in linear SVR by which controlling the hardness or softness of the margin or how much datapoints inside the margin are penalised. ϵ meanwhile defines the margin of tolerance for errors such that a higher epsilon indicates a wider margin of tolerance and a lower epsilon indicates a narrower margin of tolerance.
Since we do not have an initial range of C and ϵ values that might be suitable, we cross-validate for parameters C and ϵ by exploring model performance over a broad range of values for C and ϵ and selecting the parameters that yield the lowest MSE or MAE.

SVM with Polynomial kernel: SVM with a polynomial kernel is a technique that uses a polynomial function to calculate the similarity between different datapoints in the feature space. The polynomial function is defined as follows:

$$K(x, x') = (x \cdot x' + r)^d \tag{7}$$

Where x and x' are different input datapoints on the plane, r is an additional term, C is the cost and d is the degree. Since we do not have an initial range of C and degree values that might be suitable, we cross-validate for parameters C and degree by exploring model performance over a broad range of values for C and degree and selecting the parameters that yield the lowest MSE or MAE.

GAMs: GAMs have a number of parameters that can be fine-tuned for each variable such that accuracy are maximised. The approach chosen towards parameter selection for GAMs was as follows:

- Numeric predictor variables such as *Age*, *IDS*, *BAI*, *FQ*, *LCImax* and *aedu* were smoothed using a smoothing function.
- The method "REML" or restricted maximum likelihood was chosen due to its stability and robustness over other methods such as GCV. Furthermore, REML balances the trade-off between bias and variance by preventing instances of overfitting and underfitting much better.
- From the graphs generated from the Pearson correlation test (as seen in Appendix 1) it seemed that the relationships between most numeric variables and the predictor variable were non-linear

but simple and thus did not warrant a high number of knots. Thus it was decided to experiment with a lower number of knots k=3, k=5 and k = 7.

- Lastly, thin plate splines were preferred over other kinds of splines due to their flexibility in knot placement. Furthermore, thin plate splines also include a penalty term that penalise overly fit splines, preventing over-fitting.

GBMs: Gradient Boosting methods have a number of parameters that can be fine-tuned such that accuracy are maximised. For the GBM implementation on the MH_dat dataset the distribution method chosen was Gaussian as this method is most suitable for a regression problem where the outcome is a continuous numeric variable. More intricate parameters such as the number of initial trees, the interaction depth and shrinkage were cross-validated such that the most favorable combination (the one that yielded the lowest MSE and MAE) was chosen.

Question 3

In order to interpret this dataset correctly it is also vital to understand which variables stand out as strong predictors of dep_sev_fu and which variables pose less significant impact on dep_sev_fu.
In Support Vector Machines in R we can retrieve the importance of each predictor variable by extracting its coefficients (or weights), multiply the tranpose with the support vectors and then doing an absolute of the result (to control for negative and positive weights). These values are then summed for each feature to give the importance of each feature.
The figures below show the relative importance of features in the SVRs and their direction.

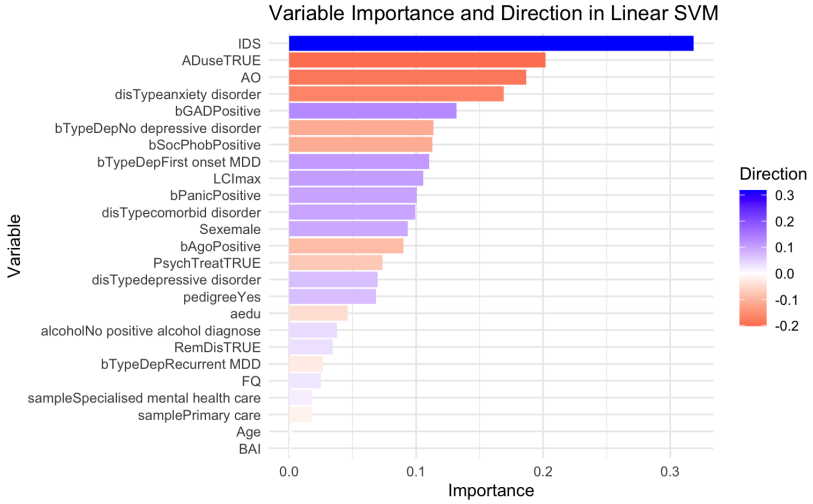

Figure 2: The figure shows the variable importance and direction of the linear SVR in the MH_dat dataset.

As seen above, in linear SVRs, the top three most important features with the most influence on dep_sev_fu are IDS which is the test score on the Inventory of Depressive Symptomatology, $ADuse$ is a variable indicating if the subject uses anti-depressant medication and AO which is a variable indicating the age at the onset of the disorder. While IDS has a positive effect on dep_sev_fu, meaning that an increase in the test score increases depression severity at follow up, ADuse and AO have a negative effect. This means that if the subject is using anti-depressant medication and has a higher age at the onset of the disorder, the depression severity is likely to be lesser.

In SVRS with radial basis kernels, the top three most important features are AO, $ADuse$ and $bTypeDepFirst$. Similar to in linear SVRs both AO and ADuse have negative effects. bTypeDepFirst on the other hand has a positive effect, indicating that if the particular category of depression is onset Major Depressive Disorder (MDD) the depression severity at follow up is likely to be higher.

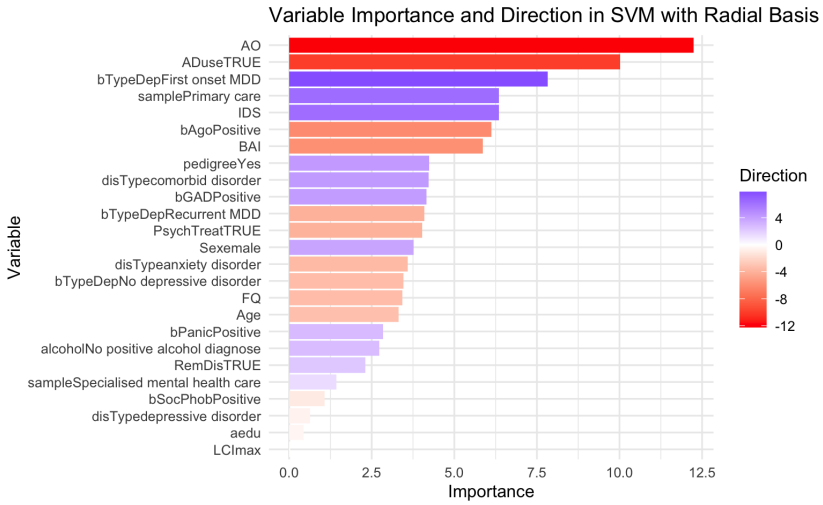

Figure 3: The figure shows the variable importance and direction of the radial SVR in the MH_dat dataset.

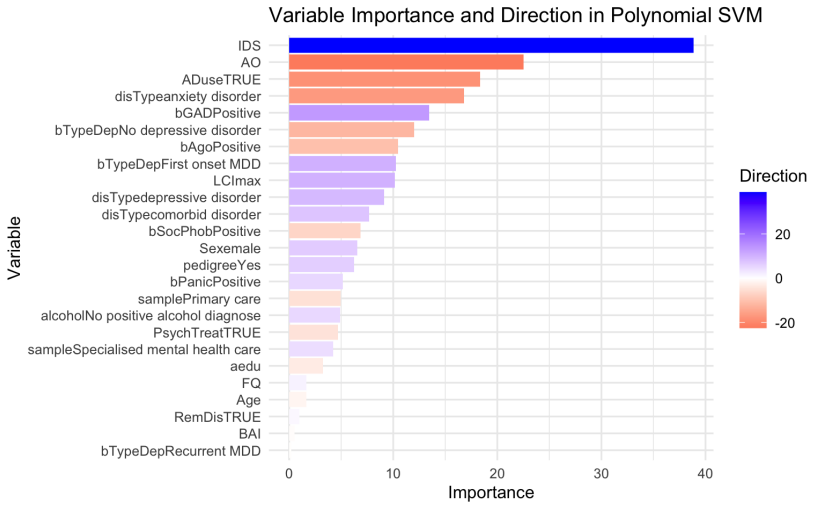

Figure 4: The figure shows the variable importance and direction of the polynomial SVR in the MH_dat dataset.

In SVRs with polynomial kernels, the top three most important features are *IDS*, *AO* and *ADuse*. Similar to in linear SVRs both AO and ADuse have negative effects and *IDS* has a positive effect.
In GAMs, the importance of variables can be getting the coefficient estimates and the direction of the influence can be determined by looking at the signs of coefficient estimates. As seen in figure 5 the most important features are *distype*, *ADuse* and *bTypeDep*. The variable *distype* indicates the specific type of disorder and from the graph we see that a diagnosis of depressive disorder or comorbidity (co-occurring mental disorders) corresponds to higher dep_sev_fu or depression severity at follow up. Furthermore, the variable *bTypeDep* indicates the particular subtype of depression. It is clear that a depression subtype of 'No depressive disorder' corresponds to lower depression severity at follow up.

For GBMs, importance of variables can be acquired by using a **VarImp()** function in R which calculates the contributions of each variable to model performance by sequentially removing each variable and seeing how the model performance is affected. The plot in figure 6 show the ranked importance of variables. As we can see from the graph *IDS*, *AO* and *disType* disorder are the top 3 variables with the highest impact on the variable dep_sev_fu. Plotting the partial plots for each variable help us understand whether the effect of these variables is positive or negative.

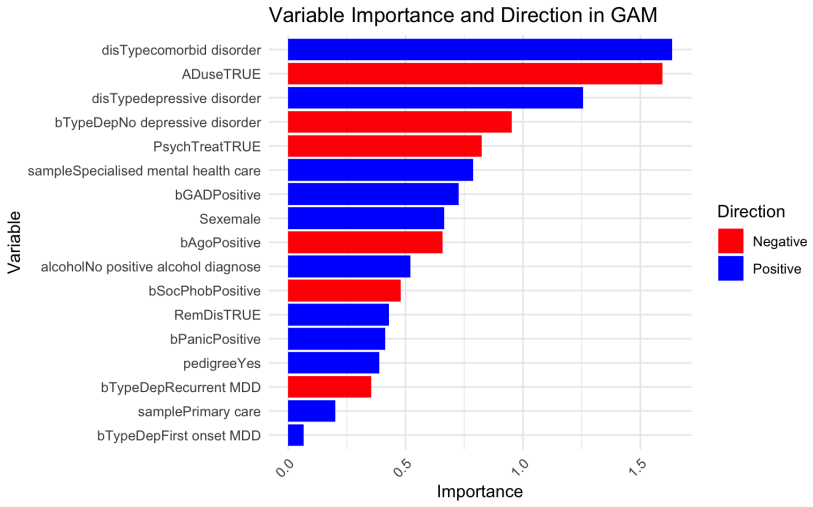

Figure 5: The figure shows the variable importance and direction of the linear SVR in the MH_dat dataset.

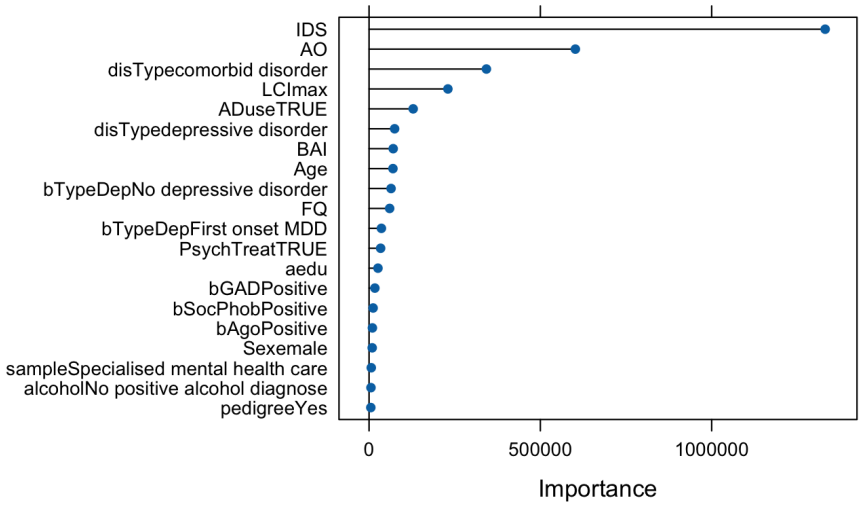

Figure 6: The figure shows the variable importance of variables in Gradient Boosted Models in the MH_dat dataset.

As seen in figures 7, 8 and 9, *IDS* has a positive effect on dep_sev_fu indicated by the positive gradient of the curve. AO has a negative effect on dep_sev_fu as indicated by the negative gradient of the curve. Lastly, we can see the effect of the categorical variable *disType* on dep_sev_fu. We can explain it as dep_sev_fu is higher when the *disType* is comorbidity (multiple co-occuring mental disorders) and lower when it is depressive disorder and lowest when it is anxiety disorder.

Question 4

To assess the performance of any statistical model it is important to quantify its performance in some measurable terms. In regression problems such as this one we can find the accuracy by calculating the mean-squared error (MSE) which measures the difference between the predicted outcome and actual outcome and the mean absolute error (MAE) which measures the average absolute error between the predicted outcome and actual outcome. A table of the performance of different models is given on page 8 in order of execution.

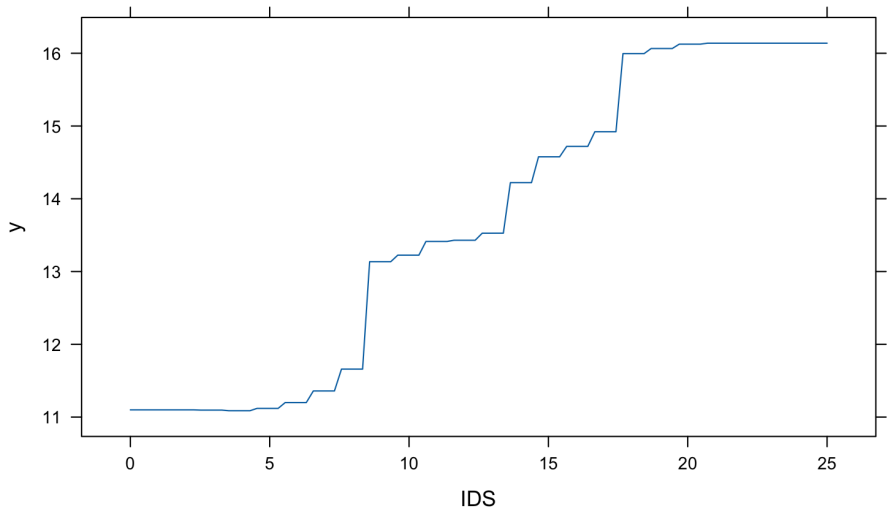

Figure 7: The figure shows the effect of IDS on dep_sev_fu variable in the MH_dat dataset.

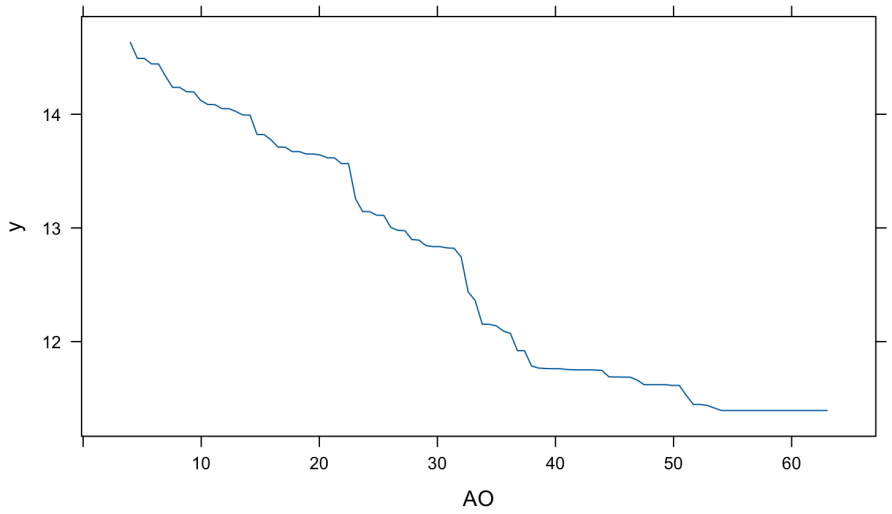

Figure 8: The figure shows the effect of AO on dep_sev_fu variable in the MH_dat dataset.

Note: The values are written to up to 3 decimal point numbers since in some cases the difference in accuracy is very small.

As seen in table 1 the models that perform the best and have the lowest MSE and MAE are Linear SVM and GBM. While many guesses can be made as to the reasons, it is obvious that GAMs although more flexible than simple linear regression models to non-linearity in data, still assume that a set of smoothing functions fitted at each predictor variable to the outcome variable is sufficient to capture the non-linearity in data when in reality it is possible that the non-linearity arises from the complexity of relationships between predictor variables themselves.

Table 2 contains the confidence intervals of the pairwise comparisons of performance between different supervised learning datasets. Since all confidence intervals contain 0 we can say that there is no significant difference in the performance of these methods.

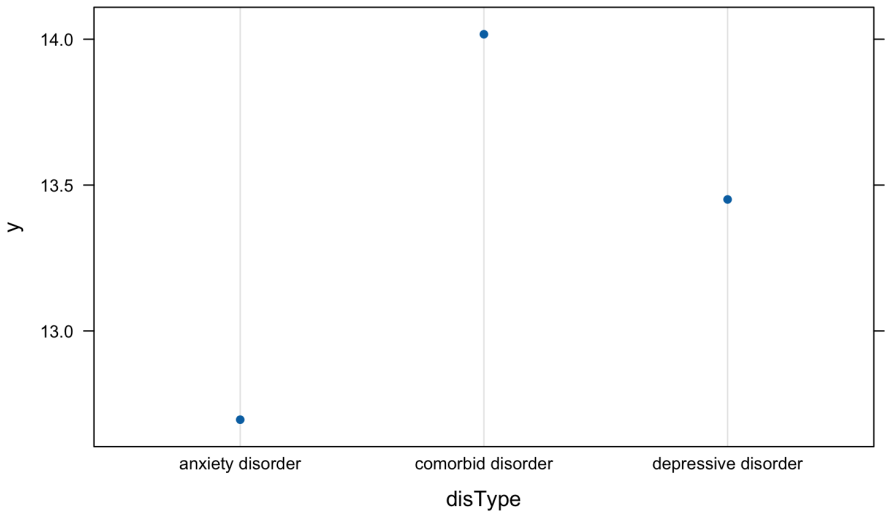

Figure 9: The figure shows the effect of *disType* on *dep_sev_fu* variable in the *MH_dat* dataset.

Model Performance		
Models	**Mean Squared Error (MSE)**	**Mean Absolute Error (MAE)**
Linear SVM	18.399	3.463
Polynomial SVM	18.929	3.514
Radial SVM	21.154	3.692
GAM with k=3	18.529	3.497
GAM with k=5	18.503	3.489
GAM with k=7	18.497	3.488
GBM*	18.399	3.463

Table 1: Performance metrics of supervised learning models on *MH_dat* dataset with * indicating where cross-validation was used to find best parameters.

Model Confidence Intervals							
Models	**Linear SVM**	**Poly. SVM**	**Radial SVM**	**GAM (k=3)**	**GAM (k=5)**	**GAM (k=7)**	**GBM***
Linear SVM	–	(-1.4157, 0.3567)	(-1.3895, 0.4380)	(-1.4320, 0.3685)	(-1.4400, 0.4425)	(-1.5455, 0.3987)	(-1.4584, 0.4002)
Poly. SVM	–	–	(-1.4343, 0.4040)	(-1.4746, 0.3677)	(-1.3676, 0.3920)	(-1.4633, 0.4023)	(-1.3880, 0.3954)
Radial SVM	–	–	–	(-1.4152, 0.4842)	(-1.3989, 0.3954)	(-1.4407, 0.4440)	(-1.5398, 0.3994)
GAM (k=3)	–	–	–	–	(-1.4307, 0.3775)	(-1.3991, 0.4055)	(-1.4798, 0.3774)
GAM (k=5)	–	–	–	–	–	(-1.4189, 0.3694)	(-1.4473, 0.4646)
GAM (k=7)	–	–	–	–	–	–	(-1.3886, 0.4616)
GBM*	–	–	–	–	–	–	–

Table 2: Confidence intervals of pairwise performative difference in supervised learning models on *MH_dat* dataset with * indicating where cross-validation was used to find best parameters.

Question 5

From the analyses of feature importance in the numerous methods implemented it is obvious that the variables *IDS*, *AO*, *disType* and *ADuse* have a significant effect on the outcome variable dep_sev_FU in the *MH_dat* dataset. This is understandable as each of these variables depicts a distinct characteristic of each individual's medical history relevant to the status of their depression or anxiety disorders. This is explained as follows:

- The variable *IDS* represents the individual's score on a 30-item self-report questionnaire to assess the severity of depressive symptoms. It is called the Inventory of Depressive Symptomatology (IDS), hence the variable name [(Rush AJ, 1986)]. Thus higher scores on this test correspond to

higher depression severity at follow up.

- The variable *ADuse* is a variable indicating whether the subject used anti-depressant medication. Higher instances of anti-depressant usage to lower depression severity at follow up and vice-versa, indicating that the medication was working.
- *AO* which is a variable indicating the age at the onset of the disorder. Higher age indicated low depression severity at follow up and vice versa. This could be because changes in depression severity are more dramatic towards the start or middle since the individual experiences more drastic developmental or lifestyle changes that could potentially exacerbate existing conditions. In contrast, individuals who experience depression at older ages are more or less protected from these drastic upheavals as they have graduated from that stage of life. Thus, this prevents their existing depression from getting worse. [(David J. Smith)]
- *disType* indicates the type of disorder. This means that an individual having a certain pre-existing disorder that exacerbates symptoms of depression such as depressive disorder or multiple such disorders is likely to have higher depression severity at follow up than individuals with ailments that do not exacerbate pre-existing depressive symptoms.

Question 6

The models' predictive performance was tested with David Edgar. The results are shown in the table below.

David Edgar's Diagnosis	
Models	dep_sev_fu
Linear SVM	16.67
Polynomial SVM	14.83
Radial SVM	15.48
GAM with k=7	17.48
GBM*	15.95

From the table above we can see the various diagnoses for David Edgar. Since all of them are below 17 we can safely say that David Edgar does not need to be referred to the intensive treatment program.

References

Douglas H. R. Blackwood David J. Smith. Depression in young adults. URL https://www.cambridge.org/core/journals/advances-in-psychiatric-treatment/article/depression-in-young-adults/B7DB64F1343880E4F19B30DA709D2FDB.

Schlesser MA Fulton CL Weissenburger J Burns C Rush AJ, Giles DE. The inventory for depressive symtomatology (ids). 1986. URL https://pubmed.ncbi.nlm.nih.gov/3737788/.

Appendix 1

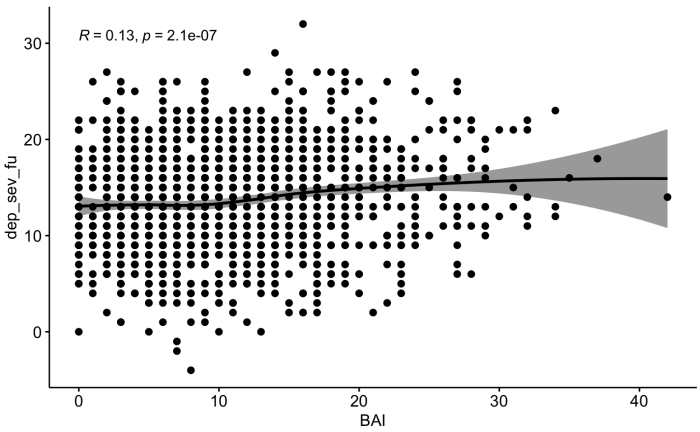

Figure 10: The figure shows the correlation between *dep_sev_fu* and *BAI*.

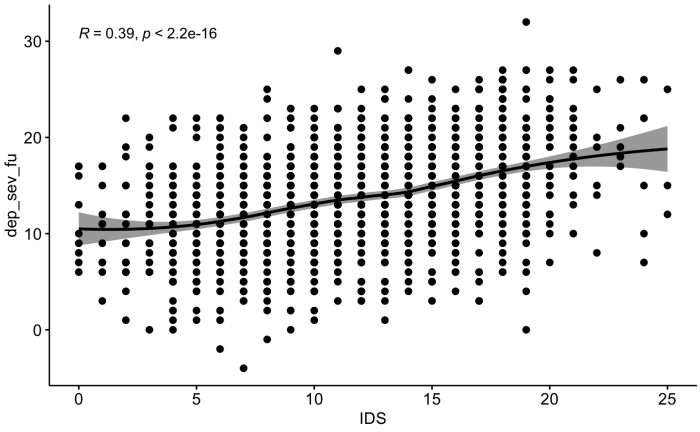

Figure 11: The figure shows the correlation between *dep_sev_fu* and *IDS*.

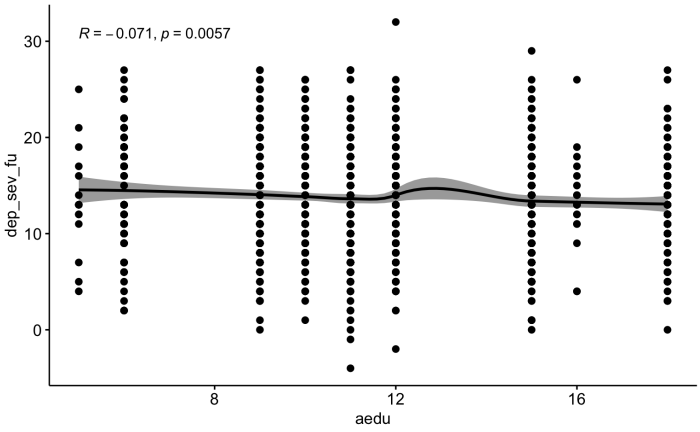

Figure 12: The figure shows the correlation between *dep_sev_fu* and *aedu*.

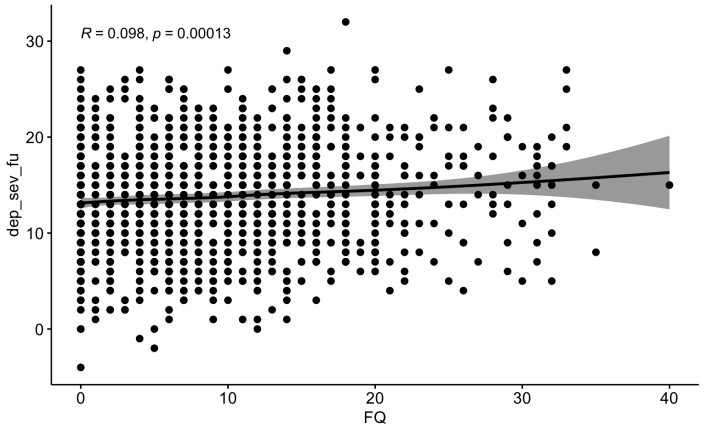

Figure 13: The figure shows the correlation between dep_sev_fu and FQ.

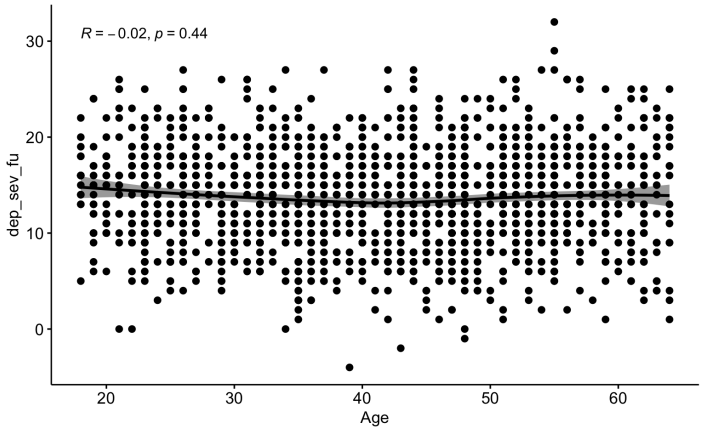

Figure 14: The figure shows the correlation between dep_sev_fu and Age.

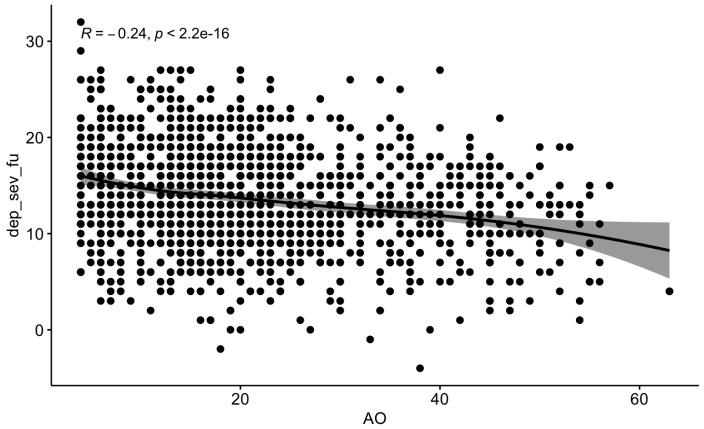

Figure 15: The figure shows the correlation between dep_sev_fu and AO.

Appendix 2

IDS	ADuseTRUE
0.3182546747	0.2016774643
AO	disTypeanxiety disorder
0.1864605897	0.1688025234
bGADPositive	bTypeDepNo depressive disorder
0.1317679065	0.1137352557
bSocPhobPositive	bTypeDepFirst onset MDD
0.1125832957	0.1103970464
LCImax	bPanicPositive
0.1056683446	0.1007054698
disTypecomorbid disorder	Sexemale
0.0990863004	0.0933542630
bAgoPositive	PsychTreatTRUE
0.0900000000	0.0736162767
disTypedepressive disorder	pedigreeYes
0.0697162230	0.0684544935
aedu	alcoholNo positive alcohol diagnose
0.0458827676	0.0376022900
RemDisTRUE	bTypeDepRecurrent MDD
0.0341888875	0.0266617907
FQ	samplePrimary care
0.0250556508	0.0178385060
sampleSpecialised mental health care	Age
0.0178385060	0.0025331759
BAI	
0.0009288078	

Figure 16: The figure shows the variable importance of the linear SVR in the MH_dat dataset.

AO	ADuseTRUE
12.23669751	10.02616795
bTypeDepFirst onset MDD	samplePrimary care
7.82808215	6.35024707
IDS	bAgoPositive
6.34204647	6.11824111
BAI	pedigreeYes
5.86101474	4.24349643
disTypecomorbid disorder	bGADPositive
4.21836998	4.16504666
bTypeDepRecurrent MDD	PsychTreatTRUE
4.09813610	4.02140179
Sexemale	disTypeanxiety disorder
3.77413558	3.58847819
bTypeDepNo depressive disorder	FQ
3.45556692	3.42678365
Age	bPanicPositive
3.31004191	2.84402086
alcoholNo positive alcohol diagnose	RemDisTRUE
2.72252630	2.31233365
sampleSpecialised mental health care	bSocPhobPositive
1.42487686	1.06937100
disTypedepressive disorder	aedu
0.62989180	0.43715907
LCImax	
0.02594711	

Figure 17: The figure shows the variable importance of the radial SVR in the MH_dat dataset.

IDS	AO
38.8373808	22.5129241
ADuseTRUE	disTypeanxiety disorder
18.3423455	16.8210555
bGADPositive	bTypeDepNo depressive disorder
13.4331572	12.0135242
bAgoPositive	bTypeDepFirst onset MDD
10.4461443	10.2444899
LCImax	disTypedepressive disorder
10.1799292	9.1271951
disTypecomorbid disorder	bSocPhobPositive
7.6938604	6.8592669
Sexemale	pedigreeYes
6.5336154	6.2242900
bPanicPositive	samplePrimary care
5.1558626	4.9597278
alcoholNo positive alcohol diagnose	PsychTreatTRUE
4.9115712	4.6694680
sampleSpecialised mental health care	aedu
4.2450533	3.2308948
FQ	Age
1.6593884	1.6423822
RemDisTRUE	BAI
0.9958754	0.5005937
bTypeDepRecurrent MDD	
0.2309657	

Figure 18: The figure shows the variable importance of the polynomial SVR in the MH_dat dataset.